等效性检验——结构方程模型评价和测量不变性分析的新视角*

王阳[1]　温忠麟[2]　付媛姝[3]

([1]广东金融学院公共管理学院，广州 510521)([2]华南师范大学心理学院/心理应用研究中心，广州 510631)([3]肇庆学院教育科学学院，肇庆 526061)

摘　要　常用的结构方程模型拟合指数存在一定局限，如χ^2以传统零假设为目标假设，无法验证模型，而 RMSEA 和 CFI 等描述性的拟合指数不具备推断统计性质，等效性检验有效弥补了这些问题。首先说明等效性检验如何评价单个模型的拟合，并解释其与零假设检验的不同，然后介绍等效性检验如何分析测量不变性，接着用实证数据展示了等效性检验在单个模型评价和测量不变性检验中的效果，并与传统模型评价方法比较。

关键词　结构方程模型；拟合指数；等效性检验；零假设检验；测量不变性

chinaXiv:202007.00045v1

1 引言

在心理学、管理学等社会科学研究领域，结构方程模型(Structural Equation Model, SEM)是一种重要的统计工具，它能够更好地控制测量误差，并同时构建复杂的多变量模型，从而帮助研究者获得比一般回归分析更加准确的分析结果。结构方程建模的一个关键步骤是评价模型拟合，因为对模型有关参数的估计要以良好模型拟合为前提。如果模型拟合不佳，即理论模型为误设模型，则建立在此模型之上的参数估计(如因子载荷、路径系数及基于模型的信度指标等)结果都是不可靠的。

尽管模型拟合至关重要，但当前通用的模型拟合评价标准尚存一些不足。已有的模型拟合指数可分两类(Marcoulides & Yuan, 2017)。一类是描述性的，最常用的是 RMSEA 和 CFI (Lai, 2020)。很多研究者通过简单报告 RMSEA 和 CFI 估计值并将其与一个经验临界值比较来支持他们的模型。这个过程不涉及显著性检验，就不知道推断错误的概率，因此很难确定怎样的拟合指数才是可以接受的(McNeish et al., 2018)；另一类是推断性的，如χ^2检验和χ^2差值检验(用于嵌套模型比较和测量不变性检验)，可以通过显著性检验来评价模型拟合好坏。然而，χ^2存在很多问题。通常的统计检验都是将想要验证的假设作为备择假设，若统计

收稿日期：2020-03-06

* 本研究得到国家自然科学基金项目(31771245)和广东省普通高校创新团队项目(人文社科)(2019WCXTD005)的资助。

通信作者：温忠麟，E-mail: wenzl@scnu.edu.cn

显著，拒绝零假设接受备择假设时知道犯错误(即第一类错误)的概率(即显著性水平)，但模型拟合χ^2检验却将想要验证的假设作为零假设，统计不显著接受零假设时犯错误(即第二类错误)的概率通常都比较大，而且还不知道有多大，χ^2检验无法真正起到验证模型的作用。

为改善上述模型拟合评价标准的不足，研究者提出将等效性检验(Equivalence Testing, ET)用于结构方程模型评价(Yuan & Chan, 2016; Yuan et al., 2016)。该方法一方面修正了χ^2检验的问题，另一方面，将推断统计元素融入了 CFI 和 RMSEA，可以用于单个模型拟合评价和测量不变性检验，已经得到了越来越多的关注和应用(如 Alpizar, 2020; Fu et al., 2018; Swami & Barron, 2019; Tóth-Király et al., 2018; Wang et al., 2020)。本文介绍等效性检验如何用于单个模型评价和测量不变性检验，然后进行实例演示，最后对相关问题进行了讨论和拓展。

2 等效性检验用于评价单个模型拟合

拟合函数 $F_{ml}[S, \Sigma(\theta)]$是结构方程模型评价的一个重要概念，它代表了样本协方差矩阵 S 与假设模型隐含的协方差矩阵$\Sigma(\theta)$的距离。应用研究者也可以简单地理解为 S 代表实际数据所反映的变量关系，$\Sigma(\theta)$代表我们假设的变量关系，拟合函数衡量了变量关系在假设模型与实际数据中的差距(吴明隆, 2010)。利用特定的算法(如极大似然估计)，可以求得一个使拟合函数 $F_{ml}[S, \Sigma(\theta)]$达到最小值的$\theta$估计值，这个拟合函数的最小值用 F_{ml} 表示。F_{ml0} 为 F_{ml} 对应的总体真值，即$\Sigma(\theta)$与真模型协方差矩阵Σ的最小距离。在很一般的条件下，随着样本容量增大，F_{ml} 收敛于 F_{ml0}(温忠麟等, 2012)。对于错误设定的模型，$F_{ml0}>0$，它的取值反映了误设大小，此时似然比统计量 T_{ml}(即前文提到的χ^2检验统计量)$=(N-1)F_{ml}$ 近似服从一个非中心化的卡方分布$\chi^2(\delta)$，其自由度 $df=p(p+1)/2-q$，p 是显变量个数，q 是自由参数个数；$\delta=(N-1)F_{ml0}$ 是非中心参数(Noncentrality Parameter)；如果模型正确设定，即 $F_{ml0}=0$，T_{ml} 服从一个中心化的卡方分布χ^2，$\delta=0$(Yuan et al., 2016)。

χ^2检验的传统零假设为：H_0：$F_{ml0}=0$ 或者 H_0：$\Sigma=\Sigma(\theta)$；而对于等效性检验，零假设变为：$H_{0\alpha}$：$F_{ml0}>\varepsilon_0$ (Yuan et al., 2016)。ε_0 是一个小正数，代表一个由研究者预先设定的可以容忍的误设程度。等效性检验的零假设也可以用非中心参数表示，即：$H_{0\alpha}$: $\delta>\delta_0$，其中$\delta_0=(N-1)\varepsilon_0$。令 $c_\alpha(\varepsilon_0)$为等效性检验的零假设分布$\chi^2(\delta_0)$在α水平(默认取 0.05)的左侧临界值，如果 $T_{ml}<c_\alpha(\varepsilon_0)$，我们拒绝零假设 $H_{0\alpha}$，认为由 F_{ml0} 代表的模型误设水平不大于ε_0，模型可以接受，做出这一推断犯错误的概率控制在α水平(Jiang et al., 2017)，也可以等价表述为有 95%的置

chinaXiv:202007.00045v1

信度认为模型可以接受。除了将 T_{ml} 和 $c_\alpha(\varepsilon_0)$相比较，我们还可以利用类似 p 值的统计量来评价等效性检验的结果：定义一个统计量ε_t，计算满足公式 $T_{ml}=c_\alpha(\varepsilon_t)$的$\varepsilon_t$，对于任何小于$\varepsilon_0$的$\varepsilon_t$，我们都可以拒绝 $H_{0\alpha}$，接受模型，并且这一推断的错误率不超过α(Marcoulides & Yuan, 2017)。这里的ε_t可以称之为最大可容忍误设(Maximum Tolerable Size of Model Misspecification)，扮演了一个类似于 p 值却又有所不同的角色。说相似是因为ε_t也通过和一个临界点比较来判断是否接受模型；说不同是因为 p 值是和显著性水平α相比，得到的结论是统计上是否显著，而ε_t是在统计显著的前提下，与一个代表误设微不足道的临界点ε_0相比较。这样，ε_t所提供的信息量比 p 值还要大，它不仅说明统计显著，还说明误设程度的效应足够小。所以更准确地说，ε_t是显著性检验和效应量的结合体。

对于等效性检验，临界点ε_0的选择对分析至关重要(Counsell et al., 2020)；因此需要有一个选择ε_0的标准。为避免选择过于随意、主观，可以通过公式$\varepsilon_0=df(\text{RMSEA}_0)^2$把$\varepsilon_0$和 RMSEA 关联起来(Yuan et al., 2016)，并选择 RMSEA_0=0.08 作为模型误设可以接受的经验临界值，从而相对客观地确定ε_0。类似地，我们也可以通过公式$\varepsilon_t=df(\text{RMSEA}_t)^2$将$\varepsilon_t$和 RMSEA 关联起来，定义一个 RMSEA_t替代ε_t的作用，它不仅可以说明模型拟合程度或误设程度大小，且推断犯错误的概率不超过α，从而使 RMSEA 具备了过去没有的统计推断性质。从 $T_{ml}=c_\alpha(\varepsilon_t)$这个方程可知，$\varepsilon_t$就是 $F_{ml0}=\delta/(N-1)$的 1-2α置信区间的上限，而 RMSEA_t是 RMSEA 的 1-2α置信区间的上限(Yuan et al., 2016)。由于 RMSEA 的 90%置信区间是 Mplus 软件的默认输出结果，在评价单个模型拟合时，我们可以直接在 Mplus 中读取 RMSEA 90%置信区间上限值，就获取了 RMSEA_t，十分方便。然后将 RMSEA_t和 RMSEA_0=0.08 相比较，如果 RMSEA_t不大于 0.08，我们可以接受模型，并声明以 RMSEA 作为模型误设评价指标时，模型误设不超过 RMSEA_t，这一推断犯错误的概率不超过 0.05；如果 RMSEA_t大于 0.08，尽管我们仍可以以 95%的置信度声明误设不超过 RMSEA_t，但因为 RMSEA_t已经高于经验临界值，无法接受模型。

3 等效性检验和传统零假设检验的比较

等效性检验和传统的零假设检验的异同点主要表现在以下几个方面(Yuan & Chan, 2016)：首先，等效性检验和零假设检验都用 T_{ml}作为统计量，但它们的零假设分布和拒绝域都不同，等效性检验的零假设假定模型存在误设(大于ε_0)，所以对应一个非中心化的卡方分布$\chi^2(\delta_0)$，当 T_{ml}落在区间$[0, c_\alpha(\varepsilon_0)]$中时，拒绝零假设，认为模型误设程度可以接受；零假设

chinaXiv:202007.00045v1

检验的零假设假定模型误设为 0，所以对应一个中心化的卡方分布χ^2，当 T_{ml} 落在区间$[c_{1-\alpha}, \infty]$中时，拒绝零假设，认为模型为误设模型。$c_{1-\alpha}$是零假设分布χ^2 在 1-α水平的右侧临界值。可见，等效性检验和零假设检验的作用刚好相反，前者适用于接受模型，而后者适用于拒绝模型。

第二，尽管名义上第一类错误的概率大小都为α，但两种检验的第一类错误的含义不同。对等效性检验而言，第一类错误是指将一个不可接受的模型(即 $F_{ml0}>\varepsilon_0$)判定为可接受的模型；而对于零假设检验，第一类错误是指将正确设定的模型(即 $F_{ml0}=0$)判定为误设模型(即 $F_{ml0}>0$)。

第三，在一定条件下，零假设检验和等效性检验的零假设可以同时被接受或拒绝，但意义不同。如果 $c_\alpha(\varepsilon_0)$小于 $c_{1-\alpha}$，且观测到的 T_{ml} 落在这两个数字之间，零假设检验和等效性检验的零假设都不能被拒绝，即既无法声明模型是无误设的，也不能证明模型误设在可接受范围内。如果 $c_\alpha(\varepsilon_0)$大于 $c_{1-\alpha}$，且观测到的 T_{ml} 落在这两个数字之间，我们可以同时拒绝零假设检验和等效性检验的零假设，即模型误设并不是 0，但在一个可接受范围内。

第四，由于 $c_\alpha(\varepsilon_0)$随着ε_0 的增加而增加，将会存在一个ε_0 使得 $c_\alpha(\varepsilon_0)$刚好等于 $c_{1-\alpha}$。此时，等效性检验可能会得到和零假设检验一样的结论。但这个结果的意义是不一样的。对于等效性检验，在 $c_\alpha(\varepsilon_0)=c_{1-\alpha}$，我们可以认为误设水平为$\varepsilon_0=F_{ml0}$ 是可以接受的；但对于零假设检验，我们不知道当前模型的误设有多大。

4 等效性检验用于测量不变性分析

测量不变性检验是结构方程模型的一种重要应用，它考察了问卷的结构在不同情景中(如不同群体和不同时间)是否是一致的。对于无均值结构的模型，传统的测量不变性分析通常可以按照如下顺序进行：(1)检验两个组有无相同的模型结构，即因子数相同和因子-条目对应关系相同，也称形态不变性(Configural Invariance)；(2)检验各组载荷是否跨组不变，即单位不变性(Metric Invariance)或弱不变性(Weak Invariance)；(3)检验各组条目误差方差是否跨组不变，即误差方差不变性(Error Variance Invariance)/严格不变性(Strict Invariance)；(4)检验各组因子方差和协方差是否跨组不变。对于有均值结构的模型，可以在单位不变性检验之后，按如下顺序进行分析：(3’)检验各组条目截距是否跨组不变，即截距不变性(Scalar Invariance)或强不变性(Strong Invariance)；(4’)检验各组潜均值是否跨组不变(Jiang et al., 2017)。

chinaXiv:202007.00045v1

每一步检验都以前一步模型成立为前提，否则停止分析。模型成立通常可以依据两个标准：其一，当前模型与前一步模型的卡方差值即$\Delta\chi^2$统计不显著；其二，当前模型与前一步模型的 CFI 或 RMSEA 差值即ΔCFI 或ΔRMSEA 足够小(Jin, 2020)。χ^2、CFI 和 RMSEA 理论上的不足前文已有说明，且已有的实证和模拟研究显示，$\Delta\chi^2$、ΔCFI 及ΔRMSEA 的统计学性质均不够理想。以错误判定不变性的概率作为第一类错误率、以正确判定不变性的概率作为统计检验力，$\Delta\chi^2$、ΔCFI 和ΔRMSEA 的第一类错误率总是明显偏高(Counsell et al., 2020; Finch & French, 2018; Yuan & Chan, 2016)，特别是样本容量偏小(如小于 250)时，最大可达 0.8 以上(Counsell et al., 2020)。ΔCFI 还有个额外的问题，在因子载荷较高时，无论样本容量多大，它的第一类错误率总是极高。在统计检验力方面，$\Delta\chi^2$ 在模型完全没有误设(即 F_{ml0}=0)时，检验力较高(Counsell et al., 2020; Finch & French, 2018; Yuan & Chan, 2016)，不过这种情况不切实际；而在更现实的条件下，即当模型存在可以忽略的误设时，它的检验力会随着样本容量增加而反常地下降(Counsell et al., 2020)。相较于这些传统方法，等效性检验第一类错误率控制很精确，基本上稳定在 0.05 水平，且不受样本容量、测量模型和因子载荷大小影响，统计检验力也可以接受(Counsell et al., 2020)。

在测量不变性分析中，等效性检验的零假设为：H_{eab}：F_{mla0}-F_{mlb0}>ε_{0ab}。ε_{0ab} 代表一个可以接受的有约束模型 a 和基准模型 b 之间的差值，其作用和单个模型评价中的ε_0类似。当$\Delta\chi^2$估计值小于其零假设分布 $\chi^2_{ab}(\delta_{0ab})$ 的左侧临界值时(δ_{0ab}=(N-m)ε_{0ab})，我们拒绝 H_{eab}，并以 1-α 的置信度声明由模型 a 的额外约束导致的误设不超过ε_{0ab}(Jiang et al., 2017)。同样地，我们也可以利用ε_t或 RMSEA$_t$代替 p 值来进行统计推断。除了测量不变性分析这种嵌套模型比较的特例，等效性检验也可以拓展应用在任何嵌套模型的比较中。

为了便于研究者通过等效性检验实现测量不变性分析，Jiang 等人(2017)编写了 R 程序包 equaltestMI。除了利用到测量学性质更好的等效性检验之外，equaltestMI 还具备两个突出优势：(1)传统的测量不变性分析中每个约束模型都要建模一次(如用 Mplus 进行不变性分析)，分析过程繁琐且易出错，而 equaltestMI 利用原始数据(包括问卷条目和分组变量)和一个简单的验证性因子分析模型设定，即可一次性输出测量不变性的每一步分析，十分方便，而且既有传统测量不变性分析的结果，也有基于等效性检验的分析结果，利于研究者对比两种方法的结果。

(2)传统的测量不变性分析中，均值结构涉及截距项和潜均值。截距项需要设为跨组不变的，从而使得潜均值可以被识别和估计。即潜均值不变性以截距不变为前提。然而，有研

chinaXiv:202007.00045v1

究者指出这在理论上是不必要的(Jiang et al., 2017)，现实中截距不变也很难满足(王孟成, 2014; Swami & Barron, 2019)。equaltestMI 利用 Deng 和 Yuan (2016)提出的投影法(Projection Method)解决了这个问题。投影法将每个组的显变量均值分解为两个正交的成分：公分数(Common Scores)和特殊因子(Specific Factors)。检验公分数均分的跨组不变性本质上是检验潜均值的跨组不变性，而检验特殊因子均分的跨组不变性和截距不变性有关，但不一样。公因子和特殊因子均分仅依赖于样本均值和估计出的公因子载荷矩阵，不会涉及到截距项，所以使用投影法就可以在不要求显变量截距不变的前提下比较潜均值，只需要满足载荷跨组不变即可，简化了测量潜均值不变性分析的步骤。更详细的原理介绍请见附录或 Deng 和 Yuan (2016)。注意：投影法既可以用于等效性检验，也可以用于传统测量不变性分析。

5 应用实例

下面用一个例子演示如何利用等效性检验评价单个模型拟合及测量不变性。示例样本为 856 名大学生，男 487 人，女 369 人。统计软件采用 Mplus 8.3 和 R 3.5.1 的 equaltestMI 包(语句示例见附录)。所有被调查者完成由王阳等人(2017)修订的共情量表，该量表共 8 个条目，包含 2 个维度：认知共情(理解他人情绪)和情绪共情(体验到他人情绪; Wang et al., 2019)。每个维度有 4 个条目。首先评价共情量表的两因子模型。传统的拟合指数 RMSEA=0.046。等效性检验结果显示 $RMSEA_t$=0.061。虽然两种方法看起来都说明模型可以接受(如果以 0.08 为临界值)，但对于前者，我们无法获取有关推断犯错误概率的信息，不知道是否模型是正确设定的，如果存在误设，也不知道误设的程度(Marcoulides & Yuan, 2017)；而等效性检验不仅告诉我们，当以 RMSEA 衡量模型误设时，模型误设程度不超过 0.061，按照传统临界值，模型的拟合可以接受。还告诉我们做出这一推断犯错误的概率不超过 5%。

接下来进行跨性别测量不变性分析，由于很多时候研究者对误差方差的不变性不感兴趣(王孟成, 2014, Svetina et al., 2020; Swami & Barron, 2019)，我们按照形态不变性→单位不变性→截距不变性→潜均值不变性的顺序进行分析(结果见表 1)。对于形态不变性，单独用男生和女生样本拟合数据，结果 RMSEA 和 CFI 都可以接受，传统方法支持形态不变性；$RMSEA_t$=0.068、0.073，表明对于男生和女生样本，我们有 95%的置信度认为模型误设分别不超过 0.068 和 0.073，等效性检验支持形态不变性。对于单位不变性和截距不变性模型，传统方法的ΔRMSEA 和ΔCFI 均小于 0.01，$\Delta\chi^2$ 也不显著，传统方法支持单位不变性和截距不变性；单位不变性和截距不变性对应的 $RMSEA_t$ 分别为 0.074 和 0.043，我们有 95%的置

chinaXiv:202007.00045v1

信度认为施加载荷相等限制和施加截距相等限制造成的误设程度各自不超过 0.074 和 0.043，等效性检验也支持单位不变性和截距不变性。对于潜均值不变性，ΔRMSEA 和ΔCFI 均小于 0.01，$\Delta\chi^2$ 也不显著，传统方法支持潜均值不变性；而 $RMSEA_t$=0.130，我们有 95%的置信度认为施加潜均值相等限制造成的误设程度不超过 0.130，这个值已经超过拟合可以接受的经验临界值 0.08(温忠麟等, 2018)，这样等效性检验未支持潜均值不变性，除非我们可以接受最大为 0.130 的误设。

表 1 测量不变性分析各项拟合指数

模型	ε_t	$RMSEA_t$	RMSEA	CFI	ΔRMSEA	ΔCFI	$\Delta\chi^2$	Δdf	$p(\Delta\chi^2)$
形态不变性(男)	0.044	0.068	0.042	0.991					
形态不变性(女)	0.051	0.073	0.055	0.981					
形态不变性			0.048	0.987					
单位不变性	0.017	0.074	0.046	0.986	0.002	0.001	8.352	6	0.213
截距不变性	0.006	0.043	0.042	0.987	0.004	0.001	3.439	6	0.752
潜均值不变性	0.017	0.130	0.043	0.986	0.001	0.001	5.373	2	0.068

由于截距不变性条件苛刻，经常不能被满足(Svetina et al., 2020)，这导致研究者更感兴趣的潜均值不变性分析无法进行。如果研究者想要跳过截距相等限制的模型，在载荷不变成立之后直接检验潜均值是否跨组不变，可以使用投影法(操作方法见附录)。本例结果表明：基于投影法的传统测量不变性分析中，公分数跨组不变模型对应的$\Delta\chi^2$不显著(p=0.508)，结果支持潜均值不变。基于投影法的等效性检验中，公分数跨组不变模型对应的 $RMSEA_t$=0.129，即我们有 95%的置信度认为，令公分数跨组不变造成的误设不超过 0.129，由于这个上限已经超过了经验临界值，因此潜均值不变性未得到支持。

上述多个结果均显示传统方法和等效性检验对潜均值是否不变结论矛盾，我们可以通过检验两组潜均值差异是否统计显著来验证哪种结果更合理。结果无论是认知共情还是情绪共情因子，潜均值差异(标准化解分别为 0.167 和 0.166)都统计显著(ps<0.05)，女生得分高于男生，这也符合相关理论和实证研究的结论(如 颜志强, 苏彦捷, 2018; Wang et al., 2017)。这说明潜均值相等实际上不成立，等效性检验给出了正确的判断，而传统方法结果错误。此外，等效性检验不变性分析的每一步都能告诉研究者可能的误设程度和统计推断犯错误的概率，而传统方法则缺乏信息量。综合以上结果，本文的应用实例也显示出相比于传统的模型评价方法，等效性检验更具优势。

chinaXiv:202007.00045v1

6 讨论

本文从结构方程传统拟合指数的不足入手，详细介绍了等效性检验这种新的拟合评价方法，并用实际数据演示了如何用等效性检验进行单个模型拟合评价和测量不变性检验，说明了等效性检验相对于传统方法的优势(提供更多的信息和更准确的结果)。但本文仍有一些相关问题需要深入讨论和拓展。

6.1 等效性检验应用于其它拟合指数

前文主要介绍了如何以 RMSEA 的形式定义误设程度。除此之外，用类似的办法也可以以其它常用的拟合指数来定义误设。如 Yuan 等人(2016)提出用公式 $T_{ml}=c_{a/2}(\varepsilon_t)$和 $T_{mli}=c_{a/2}(\varepsilon_{it})$来计算$\varepsilon_t$和$\varepsilon_{it}$(下标 i 表示独立模型)，进而求得 $CFI_t=1-(\varepsilon_t/\varepsilon_{it})$。传统的 CFI 估计值不能告诉我们它偏离总体 CFI 值多远，而 CFI_t 可以。比如，前文应用实例中，共情两因子模型的 CFI_t=0.975，这一结果说明我们有 95%的置信度认为 CFI 真值高于 0.975。目前 CFI_t 可以通过 Marcoulides 和 Yuan (2017)编写的 R 函数计算，只需要输入样本容量、变量个数、理论模型的χ^2统计量和自由度、独立模型的χ^2统计量、以及α取值即可(语句下载网址：http://www3.nd.edu/~kyuan/EquivalenceTesting/T-size_RMSEA_CFI.R)。

除了 RMSEA 和 CFI，SRMR 也是一种广受推荐的拟合指数，不过，由于该指数并不基于卡方或非中心参数，且其分布是未知的(Kelloway, 2015)，不容易像 RMSEA 和 CFI 那样改造成等效性检验的形式。将等效性检验的思想融入 SRMR 是未来研究可以尝试的工作。

6.2 拟合临界值

前文应用实例中评价拟合水平所对照的临界值仍然是传统临界值，即 RMSEA 不高于 0.08 和 CFI 不低于 0.90 表示模型拟合可以接受。Yuan 等人(2016)认为如果等效性检验仍然采用这个标准，会过于严格，$RMSEA_t$ 和 CFI_t 需要新的临界值标准。他们通过十余个和 RMSEA、CFI 有关的变量建立回归方程来预测 $RMSEA_t$ 和 CFI_t。以测定系数 R^2 最大为标准确定最佳预测变量，并将因变量预测值作为 $RMSEA_t$ 和 CFI_t 的校正临界值，用 $RMSEA_e$ 和 CFI_e 表示。校正临界值比传统的标准宽松，会随着组数、样本容量和自由度的变化而变化(Finch & French, 2018)。新的临界值可以利用 Marcoulides 和 Yuan(2017)提供的 R 函数获得(下载网址：http://www3.nd.edu/~kyuan/EquivalenceTesting/CFI_e.R 和 http://www3.nd.edu/~kyuan/EquivalenceTesting/RMSEA_e.R)，$RMSEA_e$ 也可以利用 equaltestMI 包的 adjRMSEA = TRUE 命令获得。

尽管新临界值得到一些研究者的支持(如 Finch & French, 2018; Marcoulides & Yuan,

chinaXiv:202007.00045v1

chinaXiv:202007.00045v1

2017; Yuan et al., 2016)，但要注意 $RMSEA_e$ 和 CFI_e 的提出并没有充分的理论根据，同样是带有主观性的临界值，而且 CFI_e 的计算公式的预测误差最大接近 0.03(Yuan et al., 2016)。此外，Counsell 等人(2020)最近的模拟研究显示，在进行测量不变性分析时，对照传统拟合临界值的等效性检验可以较精确地控制第一类错误，且统计检验力也在合理范围，而使用校正临界值的等效性检验的第一类错误率总是偏高，且在样本容量低于 2000 时，几乎总是 10 倍于使用传统临界值的等效性检验。因此，尽管多篇文章推荐使用，本文仍建议应用研究者慎用校正临界值。

6.3 结合多种指标评价模型拟合

尽管等效性检验在评价结构方程模型拟合和测量不变性时展现出了优于对应的传统拟合指数(如 $RMSEA_t$ 对应于 RMSEA)的性质，Yuan 等人(2016)还建议将等效性检验作为评价模型拟合指数的惯例报告内容，但其并非没有局限。首先，等效性检验通常以 RMSEA 的形式衡量误设，而 RMSEA 本身就有一定局限。如当自由度小、样本容量小或测量条目较多(如 30 以上)时，RMSEA 容易拒绝正确的模型(Shi et al., 2019)，有研究甚至建议样本容量小于 200 时不要报告 RMSEA(Taasoobshirazi & Wang, 2016)。考虑到不同的拟合指数实际上是从不同的角度反应了模型拟合，如 RMSEA 是基于卡方的绝对拟合指数，CFI 评价了假设模型比独立模型改善的程度，而 SRMR 则是少有的直接基于残差的拟合指数，同时报告多个拟合指数应该比单纯依赖于某个拟合指数更能够有效地反映模型拟合(温涵, 梁韵斯, 2015)，尽管某些拟合指数当前还无法与等效性检验结合起来。未来的研究也应该进一步比较等效性检验与这些未关联等效性检验的拟合指数，如比较 $RMSEA_t$ 和 SRMR，从而为研究者选择合适的拟合指数评价模型提供更多指导。

第二，单纯依靠拟合指数评价模型是不够的，特别是决定拟合指数取值的因素很多，拟合指数所反映的并不仅仅是模型拟合(Moshagen & Auerswald, 2018)。比如所谓的信度悖论(Reliability Paradox; Hancock & Mueller, 2011)——其它条件不变时，因子载荷越低(信度越低)的模型常常拟合越好，因子载荷越高(信度越高)的模型常常拟合越差——就体现了拟合指数反应模型质量的片面性。因此，尽管等效性检验优化了模型拟合指数，我们在评价模型时，仍然应当综合考虑多种指标，如检查路径系数、检视载荷或 R^2 、检视修正指数、检视残差矩阵等(温忠麟, 侯杰泰, 2008; 吴明隆, 2010)。

6.4 用等效性检验灵活检验测量不变性

虽然使用 equaltestMI 评价测量不变性非常便利，可以一步完成所有不变性检验步骤，但这样的设定一定程度上牺牲了测量不变性检验的灵活性。比如，在多组比较时，有些跨组

不变很难完全满足(包括截距不变性、误差方差不变性等)，这时可以考虑建立部分跨组不变(Partial Invariance)的模型，即允许个别不满足不变性标准的参数自由估计，仅使用满足不变性要求的指标进行后续更严格的检验。equaltestMI 无法建立这样的模型，此时可以先通过 Mplus 建立部分跨组不变的模型，获得当前模型与上一步模型的χ^2及自由度差值，加上组数、每组样本容量及α水平等参数，可以代入 Yuan 和 Chan(2016)提供的函数(见附录)检验不变性。

6.5 其它需要讨论的问题

(1)本文从统计原理的角度说明了等效性检验在评价单个模型拟合以及测量不变性分析时的优势，也列举了一些有关测量不变性分析的模拟研究和实证研究证据(包括本文中的应用实例演示)。不过，目前还没有很多研究比较等效性检验和传统的拟合指数在评价单个模型时的实际表现。只有一项基于单个模拟数据集(真模型为两因子模型)的研究(Marcoulides & Yuan, 2017)，结果发现传统拟合指数会错误地将单因子模型当做拟合良好的模型，而等效性检验则发现了单因子模型拟合一般。未来还是需要模拟研究来系统地比较两种方法评价单个模型拟合的优劣。

(2)当前等效性检验在结构方程领域的应用均基于正态性假设(Yuan et al., 2016)，如果数据不服从多元正态假设，等效性检验表现如何尚未可知，需要进一步研究。对总体服从正态分布的数据进行正态转换或采用 bootstrap 方法是可以考虑的思路。

(3)本文评介了等效性检验在结构方程领域的应用，除了结构方程建模，等效性检验也可以应用在其它以接受传统零假设为目标的分析当中，如各组均值无差异(Lakens, 2017)、变量之间无相关(Shiskina et al., 2018)、方差齐性检验(Kim & Cribbie, 2017)等，并表现出了良好的统计性质。相信随着更多基于等效性检验的实证和方法研究的出现，等效性检验会有越来越多的应用。

参考文献

王孟成. (2014). *潜变量建模与 Mplus 应用(基础篇)*. 重庆大学出版社.

王阳, 王才康, 温忠麟, 肖婉婷. (2017). 共情和同情量表在中国幼儿教师样本中的效度和信度. *中国临床心理学杂志, 25*(6) , 1027–1030.

温涵, 梁韵斯. (2015). 结构方程模型常用拟合指数检验的实质. *心理科学, 38*(4), 987–994.

温忠麟, 侯杰泰. (2008). 检验的临界值: 真伪差距多大才能辨别?——评《不同条件下拟合指数的表现及临界值的选择》. *心理学报. 40*(1), 119–124.

chinaXiv:202007.00045v1

温忠麟, 黄彬彬, 汤丹丹. (2018). 问卷数据建模前传. *心理科学, 41*(1), 204–210.

温忠麟, 刘红云, 侯杰泰. (2012). *调节效应和中介效应分析.* 教育科学出版社.

吴明隆. (2010). 结构方程模型——AMOS 的操作与应用(第 2 版). 重庆大学出版社.

颜志强, 苏彦捷. (2018). 共情的性别差异: 来自元分析的证据. *心理发展与教育, 34*(2), 129–136.

Alpizar, D., French, B. F., & Vo, T. T. (2020). Equivalence testing of a youth risk and needs assessment. *Journal of Psychoeducational Assessment.* Advance online publication.

Counsell, A., Cribbie, R. A., & Flora, D. B. (2020). Evaluating equivalence testing methods for measurement invariance. *Multivariate Behavioral Research, 55*(2), 312–328.

Deng, L., & Yuan, K. H. (2016). Comparing latent means without mean structure models: A projection-based approach. *Psychometrika, 81*(3), 802–829.

Finch, W. H., & French, B. F. (2018). A simulation investigation of the performance of invariance assessment using equivalence testing procedures. *Structural Equation Modeling: A Multidisciplinary Journal, 25*(5), 673–686.

Fu, Y., Wen, Z., & Wang, Y. (2018). The total score with maximal reliability and maximal criterion validity: An illustration using a career satisfaction measure. *Educational and Psychological Measurement, 78*(6), 1108–1122.

Hancock, G. R., & Mueller, R. O. (2011). The reliability paradox in assessing structural relations within covariance structure models. *Educational and Psychological Measurement, 71*(2), 306–324.

Jiang, G., Mai, Y., & Yuan, K.-H. (2017). Advances in measurement invariance and mean comparison of latent variables: Equivalence testing and a projection-based approach. *Frontiers in Psychology, 8,* Article 1823.

Jin, Y. (2020). A note on the cutoff values of alternative fit indices to evaluate measurement invariance for ESEM models. *International Journal of Behavioral Development, 44*(2), 166–174.

Kelloway, E. K. (2015). *Using Mplus for structural equation modeling: A researcher's guide* (2nd ed.). SAGE Publications.

Kim, Y. J., & Cribbie, R. A. (2018). ANOVA and the variance homogeneity assumption: Exploring a better gatekeeper. *British Journal of Mathematical and Statistical Psychology, 71*(1), 1–12.

Lai, K. (2020). Confidence interval for RMSEA or CFI difference between nonnested models. *Structural Equation Modeling: A Multidisciplinary Journal, 27*(1), 16–32.

Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science, 8*(4), 355–362.

chinaXiv:202007.00045v1

Marcoulides, K. M., & Yuan, K.-H. (2017). New ways to evaluate goodness of fit: A note on using equivalence testing to assess structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal, 24*(1), 148–153.

McNeish, D., An, J., & Hancock, G. R. (2018). The thorny relation between measurement quality and fit index cutoffs in latent variable models. *Journal of Personality Assessment, 100*(1), 43–52.

Moshagen, M., & Auerswald, M. (2018). On congruence and incongruence of measures of fit in structural equation modeling. *Psychological Methods, 23*(2), 318–336.

Shi, D., Lee, T., & Maydeu-Olivares, A. (2019). Understanding the model size effect on SEM fit indices. *Educational and Psychological Measurement, 79*(2), 310–334.

Shiskina, T., Farmus, L., & Cribbie, R. A. (2018). Testing for a lack of relationship among categorical variables. *The Quantitative Methods for Psychology, 14*(3), 167–179.

Svetina, D., Rutkowski, L., & Rutkowski, D. (2020). Multiple-group invariance with categorical outcomes using updated guidelines: An illustration using Mplus and the lavaan/semTools packages. *Structural Equation Modeling: A Multidisciplinary Journal*, 27(1), 111–130.

Swami, V., & Barron, D. (2019). Translation and validation of body image instruments: Challenges, good practice guidelines, and reporting recommendations for test adaptation. *Body Image, 31,* 204–220.

Taasoobshirazi, G., & Wang, S. (2016). The performance of the SRMR, RMSEA, CFI, and TLI: An examination of sample size, path size, and degrees of freedom. *Journal of Applied Quantitative Methods, 11*(3), 31–40.

Tóth-Király, I., Morin, A. J., Bőthe, B., Orosz, G., & Rigó, A. (2018). Investigating the multidimensionality of need fulfillment: A bifactor exploratory structural equation modeling representation. *Structural Equation Modeling: A Multidisciplinary Journal, 25*(2), 267–286.

Wang, Y., Li, Y., Xiao, W., Fu, Y., & Jie, J. (2020) Investigation on the rationality of the extant ways of scoring the Interpersonal Reactivity Index based on confirmatory factor analysis. *Frontiers in Psychology, 11,* Article 1086.

Wang, Y., Su, Q., & Wen, Z. (2019). Exploring latent profiles of empathy among Chinese preschool teachers: A person-centered approach. *Journal of Psychoeducational Assessment, 37*(6), 706–717.

Wang, Y., Wen, Z., Fu, Y., & Zheng, L. (2017). Psychometric properties of a Chinese version of the Measure of Empathy and Sympathy. *Personality and Individual Differences, 119,* 168–174.

Yuan, K.-H., & Chan, W. (2016). Measurement invariance via multigroup SEM: Issues and solutions with chi-square-difference tests. *Psychological Methods, 21*(3), 405–426.

chinaXiv:202007.00045v1

Yuan, K.-H., Chan, W., Marcoulides, G. A., & Bentler, P. M. (2016). Assessing structural equation models by equivalence testing with adjusted fit indexes. *Structural Equation Modeling: A Multidisciplinary Journal, 23*(3), 319–330.

chinaXiv:202007.00045v1

附录：

投影法原理介绍

在传统的测量不变性检验中，均值结构涉及截距项和潜均值。截距项$\boldsymbol{\gamma}^{(j)}$需要设为跨组不变的，从而使得潜均值$\boldsymbol{\tau}^{(j)}=E(\boldsymbol{\xi}^{(j)})$可以被识别和估计($\boldsymbol{\xi}^{(j)}$是 k 个因子组成的向量)。为了回避这个假设，Deng 和 Yuan(2016)提出将观测变量分解为公分数、特殊因子和测量误差。公式如下：$\mathbf{x}^{(j)}=\Lambda\mathbf{f}^{(j)}+\mathbf{u}^{(j)}+\mathbf{e}^{(j)}, j=1,...,m$，$\Lambda\mathbf{f}^{(j)}$是 p 个公分数组成的向量，$\mathbf{u}^{(j)}$是 p 个特殊因子组成的向量，$\mathbf{e}^{(j)}$是 p 个测量误差组成的向量，$E[\mathbf{f}^{(j)}]=\boldsymbol{\kappa}^{(j)}$，$E[\mathbf{u}^{(j)}]=\boldsymbol{\nu}^{(j)}$，$E[\mathbf{e}^{(j)}]=\mathbf{0}$。上述公式形式上看起来很像传统的潜均值不变性公式 $\mathbf{x}^{(j)}=\boldsymbol{\gamma}^{(j)}+\Lambda\mathbf{f}^{(j)}+\boldsymbol{\zeta}^{(j)}, j=1,...,m$，但实际上存在差异。首先，对于传统的测量不变性公式，$\boldsymbol{\zeta}^{(j)}$是特殊因子和测量误差的混合向量，而对于投影法，特殊因子和测量误差是分离开来的，$\boldsymbol{\zeta}^{(j)}=\mathbf{u}^{(j)}+\mathbf{e}^{(j)}$；第二，也是更重要的一点，传统的潜均值不变性的公式有截距项，而投影法的公式当中没有截距项，所以用投影法评价潜均值不变性不需要以截距跨组不变为前提。

假设公分数空间和特殊因子空间正交，因子 $\mathbf{f}^{(j)}$的公分数均值间的比较是独立于特殊因子 $\mathbf{u}^{(j)}$的。在投影法中，$\mathbf{x}^{(j)}$的均值结构可以分解为：$\boldsymbol{\mu}^{(j)}=\boldsymbol{\mu}_{\kappa}^{(j)}+\boldsymbol{\nu}^{(j)}$，$\boldsymbol{\mu}_{\kappa}^{(j)}=\Lambda\boldsymbol{\kappa}^{(j)}$是$\boldsymbol{\mu}^{(j)}=E(\mathbf{x}^{(j)})$投影在公因子空间的部分，而 $\boldsymbol{\nu}^{(j)}$是$\boldsymbol{\mu}^{(j)}$投影在特殊因子空间的部分。令 $\hat{\Lambda}$ 为估计出的因子载荷矩阵，$\bar{\mathbf{x}}^{(j)}$ 是第 j 个组的样本均值。通过将 $\bar{\mathbf{x}}^{(j)}$ 投影到 $\hat{\Lambda}$ 的列空间可以获得公因子均分估计值，将其标记为 $\hat{\boldsymbol{\mu}}_{\kappa}^{(j)}$。相似地，通过将 $\bar{\mathbf{x}}^{(j)}$ 投影到与 $\hat{\Lambda}$ 正交的空间，可以获得特殊因子估计均值，将其标记为 $\hat{\boldsymbol{\nu}}^{(j)}$，则有 $\bar{\mathbf{x}}^{(j)}=\hat{\boldsymbol{\mu}}_{\kappa}^{(j)}+\hat{\boldsymbol{\nu}}^{(j)}$。公因子和特殊因子均分仅依赖于样本均值和估计出的公因子载荷矩阵，不会涉及到截距项的估计。

基于投影法的传统潜均值测量不变性分析的零假设为 H_{κ}：$\kappa_d^{(j)}=\kappa^{(j)}-\kappa^{(1)}=\mathbf{0}$，$j=2,...,m$，$\hat{\kappa}_d^{(j)}$ 近似服从正态分布，这个假设可以用统计量 $T_{gls}=NF_{gls}$ 来检验。T_{gls} 近似服从一个自由度为 $df_{\kappa}=(m-1)k$ 的卡方分布。如果用等效性检验做投影法，则前述假设修改为 $H_{e\kappa}$：$F_{gls0}>\varepsilon_0$，F_{gls0} 是 F_{gls} 的总体值(Jiang et al., 2017)。评价 T_{gls} 显著性的临界值是 $\chi_{\kappa}^2(\delta_0)$ 的左侧累积概率，$\delta_0=N\varepsilon_0$。当 T_{gls} 小于临界值，我们拒绝 $H_{e\kappa}$。我们也可以通过 $\text{RMSEA}_0=(\varepsilon_0/df_{\kappa})^{1/2}$ 来设定ε_0，并利用 RMSEA_t 来检验 $H_{e\kappa}$。

用 R 的 equaltestMI 包分析测量不变性的语句模板

install.packages(“lavaan”)　#安装 lavaan 程序包，因为 equaltestMI 包的模型设定是基于 lavaan 的

install.packages(“equaltestMI”) #安装 equaltestMI 程序包

library(equaltestMI) #加载 equaltestMI 包

etdata<-read.csv("C:/Users/admin/Desktop/共情.CSV", header=F)

#导入数据文件，并命名为 etdata，做自己的分析需将上述路径替换为自己的数据文件路径

#header=F 表示数据文件未含变量名

names(etdata)[1:9]<-c("gender","E1","E2","E3","E4","E5","E6","E7","E8") #变量命名，相当于 Mplus 中的 variables 模块下的“names=”命令

```
Emodel<- 'CE =~ E1 + E2 + E3 + E4
          AE =~ E5 + E6 + E7 + E8'
```

#构建测量模型，模型命名为 Emodel，CE 为认知共情，AE 为情绪共情，E1~ E8 是量表条目，“=~”相当于 Mplus 中的“by”命令

```
test <- eqMI.main(model = Emodel, data = etdata,
                  group = "gender", meanstructure = TRUE,
                  output = 'both', quiet = FALSE,
                  equivalence.test = TRUE, adjRMSEA = FALSE,
                  projection = TRUE, bootstrap = FALSE)
```

#model 代入模型名，data 代入数据名，group 代入分组变量即 gender；meanstructure=TRUE 表示包含均值结构；adjRMSEA =FALSE 表示不使用校正的 RMSEA 临界值即 $RMSEA_e$，如需使用校正临界值，将 FALSE 改为 TRUE 即可；projection = TRUE 表示采用投影法分析潜均值不变性。

注：如需获取单个模型评价的校正临界值 $RMSEA_e$，可通过 equaltestMI 包的 eqMI.RMSEA 命令，如：eqMI.RMSEA(N = 856, m = 1, df = 19)。组数 m 必须指定为 1。

用 Yuan 和 Chan(2016)的 R 函数获取测量不变性分析需要的 $RMSEA_t$

alpha=.05;N_1=487; N_2=369;N=N_1+N_2;m=2;T_ml=3.439;df=6;

#输入显著性水平、各组样本容量、组数及约束模型和上一步模型的卡方差值及自由度差值，如果需要做部分不变性分析，可以先利用 Mplus 获取上述参数，然后输入此函数

ncp=ncp_chi2(alpha=alpha, T=T_ml, df=df, m=m, N=N);

#输出非中心参数，此句不要改动

RMSEA_ctoa(m=m, N_sample=N, df=df);

#输出 $RMSEA_t$，此句不要改动

chinaXiv:202007.00045v1

注：篇幅所限，上述语句只涉及需要使用者输入的参数，而省略了 ncp_chi2 函数和 RMSEA_ctoa 函数的具体内容，感兴趣的读者需在 Yuan 和 Chan(2016) 提供的网址 (http://www3.nd.edu/~kyuan/mgroup/Equivalence-testing.R)下载完整语句，方可运行程序。

Equivalence testing——A new perspective on structural equation model evaluation and measurement invariance analysis

WANG Yang[1]; WEN Zhonglin[2]; FU Yuanshu[3]

([1] *School of Public Administration, Guangdong University of Finance, Guangzhou* 510521, *China*)([2] *School of Psychology/Center for Studies of Psychological Application, South China Normal University, Guangzhou* 510631, *China*)([3] *School of Education, Zhaoqing University, Zhaoqing* 526061, *China*)

Abstract: There are some limitations in the commonly used fit indices of structural equation model. For example, χ^2, which is set up to reject the traditional null hypothesis, cannot be used to endorse a model, while descriptive fit indices such as RMSEA and CFI do not have inferential statistical properties. Equivalence testing can effectively compensate for the limitations mentioned above. In this paper, the way to use equivalence testing in evaluating the fit of a single model and its difference from the null hypothesis testing were introduced first. Then the approach to analyze measurement invariance by equivalence testing was described. Furthermore, empirical data was used to demonstrate the effectiveness of equivalence testing in single model evaluation and measurement invariance test, and to compare equivalence testing with traditional model evaluation method.

Key words: structural equation model; fit indices; equivalence testing; null hypothesis testing; measurement invariance

chinaXiv:202007.00045v1